

**APPLICATION FOR UNITED STATES PATENT
COMPUTER-AIDED NUCLEIC ACID SEQUENCING**

By Inventors: Lubert Stryer
 843 Sonoma Terrace
 Stanford, CA 94305
 Citizenship: USA

Assignee: Affymetrix, Incorporated
 3380 Central Expressway
 Santa Clara, CA 95051
 A California Company

Entity: Large

Ritter, Lang & Kaplan LLP
12930 Saratoga Ave., Suite D1
Saratoga, CA 95070
(408) 446-8690

1004953-121101

COMPUTER-AIDED NUCLEIC ACID SEQUENCING**COPYRIGHT NOTICE**

A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the xerographic reproduction by anyone of the patent document or the patent disclosure in exactly the form it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

SOFTWARE APPENDICES

Software Appendices A and B comprising six (6) sheets are included herewith.

BACKGROUND OF THE INVENTION

The present invention relates to the field of computer systems. More specifically, the present invention relates to computer systems for sequencing biological molecules including nucleic acids.

Devices and computer systems for forming and using arrays of materials on a substrate are known. For example, PCT applications WO92/10588 and 95/11995, incorporated herein by reference for all purposes, describe techniques for sequencing or sequence checking nucleic acids and other materials. Arrays for performing these operations may be formed in arrays according to the methods of, for example, the pioneering techniques disclosed in U.S. Patent Nos. 5,445,934 and 5,384,261, and U.S. Patent Application No. 08/249,188, each incorporated herein by reference for all purposes.

According to one aspect of the techniques described therein, an array of nucleic acid probes is fabricated at known locations on a chip or substrate. A labeled nucleic acid is then brought into contact with the chip and a scanner generates an image file (also called a cell file) indicating

the locations where the labeled nucleic acids are bound to the chip. Based upon the image file and identities of the probes at specific locations, it becomes possible to extract information such as the nucleotide or monomer sequence of DNA or RNA. Such systems have been used to form, for example, arrays of DNA that may be used to study and detect mutations relevant to genetic diseases, cancers, infectious diseases, HIV, and other genetic characteristics.

The VLSIPS™ technology provides methods of making very large arrays of oligonucleotide probes on very small chips. See U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated by reference for all purposes. The oligonucleotide probes on the DNA probe array are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

For sequence checking applications, the chip may be tiled for a specific target nucleic acid sequence. For example, the chip may contain probes that are perfectly complementary to the target sequence and probes that differ from the target sequence by a single base mismatch. These probes are tiled on a chip in rows and columns of cells, where each cell includes multiple copies of a particular probe. Additionally, "blank" cells may be present on the chip which do not include any probes. As the blank cells contains no probes, labeled targets should not bind specifically to the chip in this area. Thus, a blank cell provides a measure of the background intensity.

For de novo sequencing applications, the chip may include all the possible probes of a specific length. These probes are synthesized on the chip at known locations, typically with multiple copies of a particular probe in a cell. Blank cells may also be utilized to provide a measure of the background intensity.

SUMMARY OF THE INVENTION

The present invention provides an improved computer-aided system for sequencing sample nucleic acid sequences from

nucleic acid hybridization information. The accuracy of nucleic acid sequencing is increased by analyzing the hybridization strength of related probes, where the related probes are identified according to mismatch information among the probes. The related probes may include single base mismatches or otherwise have identical subsequences. The methods of the present invention allow sequencing under conditions that do not allow identification of all of the probes that are perfectly complementary to part of the target nucleic acid sequence.

According to one aspect of the present invention, a computer system is used to sequence a nucleic acid by a method including the steps of: inputting hybridization intensities for a plurality of nucleic acid probes, the nucleic acid probes hybridizing with the nucleic acid sequence under conditions that do not allow identification of all of nucleic acid probes that are perfectly complementary to part of the nucleic acid sequence; and sequencing the nucleic acid sequence according to selected nucleic acid probes.

According to another aspect of the present invention, a computer system is used to sequence a nucleic acid by a method including the steps of: inputting hybridization intensities for a plurality of nucleic acid probes; selecting nucleic acid probes with highest numbers of single base mismatch neighbors among the probes, a single base mismatch neighbor being another probe that has the same sequence except for a single base that is different; and sequencing the nucleic acid sequence according to the selected nucleic acid probes.

According to another aspect of the present invention, a computer system is used to sequence a nucleic acid by a method including the steps of: inputting hybridization intensities for a plurality of nucleic acid probes; selecting nucleic acid probes that have fewer than a predetermined number of base mismatches with another probe; and sequencing the nucleic acid sequence according to the selected nucleic acid probes.

According to another aspect of the present invention, a nucleic acid is sequenced by a method including the steps of: contacting a set of oligonucleotide probes of predetermined sequence and length with the nucleic acid under hybridization conditions that do not allow differentiation between (i) those probes of the set which are perfectly complementary to part of the nucleic acid and (ii) those probes that are not perfectly complementary to part of the nucleic acid; selecting a subset of oligonucleotide probes that includes probes that are perfectly complementary to part of the nucleic acid and probes that are not perfectly complementary to part of the nucleic acid; and determining the sequence of the nucleic acid by compiling overlapping sequences of the subset of probes.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates an example of a computer system used to execute the software of the present invention;

Fig. 2 shows a system block diagram of a typical computer system used to execute the software of the present invention;

Fig. 3 illustrates an overall system for forming and analyzing arrays of biological materials such as DNA or RNA;

Fig. 4 is an illustration of the software for the overall system;

Fig. 5 illustrates conceptually the binding of probes on chips;

Fig. 6 shows a high level flow of sequencing utilizing mismatch information;

Fig. 7 shows a high level flow of another embodiment of sequencing utilizing mismatch information;

Fig. 8 shows a straight mismatch matrix for use with the process of Fig. 7; and

Fig. 9 shows a skewed mismatch matrix for use with the process of Fig. 7.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Introduction

In the description that follows, the present invention will be described in reference to a Sun Workstation in a UNIX environment. The present invention, however, is not limited to any particular hardware or operating system environment. Instead, those skilled in the art will find that the systems and methods of the present invention may be advantageously applied to a variety of systems, including IBM personal computers running MS-DOS or Microsoft Windows. Therefore, the following description of specific systems are for purposes of illustration and not limitation.

Fig. 1 illustrates an example of a computer system used to execute the software of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a floppy disk drive 14 and a hard drive (not shown) that may be utilized to store and retrieve software programs including computer readable code incorporating the present invention. Although a floppy disk 15 is shown as the removable media, other removable tangible media including CD-ROM, flash memory and tape may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute the software of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 52, system memory 54, I/O controller 56, display adapter 58, serial port 62, disk 64, network interface 66, and speaker 68. Disk 64 is representative of an internal hard drive, floppy drive, CD-ROM, flash memory, tape, or any other storage medium. Other

computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 52 (i.e., a multi-processor system) or memory cache.

Arrows such as 70 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, speaker 68 could be connected to the other subsystems through a port or have an internal direct connection to central processor 52. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art.

The present invention provides methods of analyzing hybridization intensity files for a chip containing hybridized nucleic acid probes. In a representative embodiment, the files represent fluorescence data from a biological array, but the files may also represent other data such as radioactive intensity, light scattering, refractive index, conductivity, electroluminescence, or large molecule detection data. Therefore, the present invention is not limited to analyzing fluorescence measurements of hybridizations but may be readily utilized to analyze other measurements of hybridization.

For purposes of illustration, the present invention is described as being part of a computer system that designs a chip mask, synthesizes the probes on the chip, labels the nucleic acids, and scans the hybridized nucleic acid probes. Such a system is fully described in U.S. Patent Application No. 08/249,188 which has been incorporated by reference for all purposes. However, the present invention may be used separately from the overall system for analyzing data generated by such systems.

Fig. 3 illustrates a computerized system for forming and analyzing arrays of biological materials such as RNA or DNA. A computer 100 is used to design arrays of biological polymers such as RNA or DNA. The computer 100 may be, for example, an appropriately programmed Sun Workstation or

personal computer or workstation, such as an IBM PC equivalent, including appropriate memory and a CPU as shown in Figs. 1 and 2. The computer system 100 obtains inputs from a user regarding characteristics of a gene of interest, and other inputs regarding the desired features of the array. Optionally, the computer system may obtain information regarding a specific genetic sequence of interest from an external or internal database 102 such as GenBank. The output of the computer system 100 is a set of chip design computer files 104 in the form of, for example, a switch matrix, as described in PCT application WO 92/10092, and other associated computer files.

The chip design files are provided to a system 106 that designs the lithographic masks used in the fabrication of arrays of molecules such as DNA. The system or process 106 may include the hardware necessary to manufacture masks 110 and also the necessary computer hardware and software 108 necessary to lay the mask patterns out on the mask in an efficient manner. As with the other features in Fig. 3, such equipment may or may not be located at the same physical site, but is shown together for ease of illustration in Fig. 3. The system 106 generates masks 110 or other synthesis patterns such as chrome-on-glass masks for use in the fabrication of polymer arrays.

The masks 110, as well as selected information relating to the design of the chips from system 100, are used in a synthesis system 112. Synthesis system 112 includes the necessary hardware and software used to fabricate arrays of polymers on a substrate or chip 114. For example, synthesizer 112 includes a light source 116 and a chemical flow cell 118 on which the substrate or chip 114 is placed. Mask 110 is placed between the light source and the substrate/chip, and the two are translated relative to each other at appropriate times for deprotection of selected regions of the chip. Selected chemical reagents are directed through flow cell 118 for coupling to deprotected regions, as well as for washing and other operations. All operations are preferably directed by an appropriately programmed computer 119, which may or may

not be the same computer as the computer(s) used in mask design and mask making.

The substrates fabricated by synthesis system 112 are optionally diced into smaller chips and exposed to marked targets. The targets may or may not be complementary to one or more of the molecules on the substrate. The targets are marked with a label such as a fluorescein label (indicated by an asterisk in Fig. 3) and placed in scanning system 120. Scanning system 120 again operates under the direction of an appropriately programmed digital computer 122, which also may or may not be the same computer as the computers used in synthesis, mask making, and mask design. The scanner 120 includes a detection device 124 such as a confocal microscope or CCD (charge-coupled device) that is used to detect the location where labeled target (*) has bound to the substrate. The output of scanner 120 is an image file(s) 124 indicating, in the case of fluorescein labeled target, the fluorescence intensity (photon counts or other related measurements, such as voltage) as a function of position on the substrate. Since higher photon counts will be observed where the labeled target has bound more strongly to the array of polymers (e.g., DNA probes on the substrate), and since the monomer sequence of the polymers on the substrate is known as a function of position, it becomes possible to determine the sequence(s) of polymer(s) on the substrate that are complementary to the target.

The image file 124 is provided as input to an analysis system 126 that incorporates the visualization and analysis methods of the present invention. Again, the analysis system may be any one of a wide variety of computer system(s), but in a preferred embodiment the analysis system is based on a Sun Workstation or equivalent. The present invention provides various methods of analyzing the chip design files and the image files, providing appropriate output 128. The present invention may further be used to identify specific mutations in a target such as DNA or RNA.

Fig. 4 provides a simplified illustration of the overall software system used in the operation of one

embodiment of the invention. As shown in Fig. 4, in some cases (such as sequence checking systems) the system first identifies the genetic sequence(s) or targets that would be of interest in a particular analysis at step 202. The sequences of interest may identify a virus, microorganism or individual. Additionally, the sequence of interest may provide information about genetic diseases, cancers or infectious diseases. Sequence selection may be provided via manual input of text files or may be from external sources such as GenBank. In a preferred embodiment that performs de novo sequencing of target nucleic acids, this steps is not necessary as the chip includes all the possible n-mer probes (where n represents the length of the nucleic acid probe).

For de novo sequencing, a chip may be synthesized to include cells containing all the possible probes of a specific length. For example, a chip may be synthesized that includes all the possible 8-mer DNA probes. Such a chip would have 65,536 cells ($4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4 \times 4$), with each cell corresponding to a particular probe. A chip may also include other probes including all the probes of other lengths.

At step 204 the system determines which probes would be desirable on the chip, and provides an appropriate "layout" on the chip for the probes. The layout implements desired characteristics such as an arrangement on the chip that permits "reading" of genetic sequence and/or minimization of edge effects, ease of synthesis, and the like.

Again referring to Fig. 4, at step 206 the masks for the synthesis are designed. At step 208 the software utilizes the mask design and layout information to make the DNA or other polymer chips. This software 208 will control, among other things, relative translation of a substrate and the mask, the flow of desired reagents through a flow cell, the synthesis temperature of the flow cell, and other parameters. At step 210, another piece of software is used in scanning a chip thus synthesized and exposed to a labeled target. The software controls the scanning of the chip, and stores the data thus obtained in a file that may later be utilized to extract sequence information.

At step 212 a computer system according to the present invention utilizes the layout information and the fluorescence information to evaluate the hybridized nucleic acid probes on the chip. Among the important pieces of information obtained from DNA probe arrays are the identification of mutant targets and determination of the genetic sequence of a particular target.

Fig. 5 illustrates the binding of a particular target DNA to an array of DNA probes 114. As shown in this simple example, the following probes are formed in the array:

```

3'-AGAACGT
AGACCGT
AGAGCGT
AGATCGT
.
.
.

```

As shown, when the fluorescein-labeled (or otherwise marked) target 5'-TCCTGCA is exposed to the array, it is complementary only to the probe 3'-AGAACGT, and fluorescein will be primarily found on the surface of the chip where 3'-AGAACGT is located. The chip contains cells that include multiple copies of a particular probe. Thus, the image file will contain fluorescence intensities, one for each probe (or cell). By analyzing the fluorescence intensities associated with a specific probe, it becomes possible to extract sequence information from such arrays using the methods of the invention disclosed herein.

For ease of reference, one may call bases by assigning the bases the following codes:

<u>Code</u>	<u>Group</u>	<u>Meaning</u>
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T(U)	Thymine (Uracil)
M	A or C	aMino
R	A or G	puRine
W	A or T(U)	Weak interaction (2 H bonds)
Y	C or T(U)	pYrimidine
S	C or G	Strong interaction (3 H bonds)

K	G or T(U)	Keto
V	A, C or G	not T(U)
H	A, C or T(U)	not G
D	A, G or T(U)	not C
B	C, G or T(U)	not A
N	A, C, G, or T(U)	Insufficient intensity to call
X	A, C, G, or T(U)	Insufficient discrimination to call

Most of the codes conform to the IUPAC standard. However, code N has been redefined and code X has been added.

Sequencing Utilizing Mismatch Information

The present invention provides methods of sequencing nucleic acid sequences utilizing mismatch information. When used herein, "mismatch information" relates to base mismatches between or among nucleic acid probes. Mismatch information may include the number of base mismatches, the location of the base mismatches and the base differences. The mismatch information maybe combined with information from the hybridization intensity to sequence the nucleic acid sequence with a high degree of accuracy. In a preferred embodiment, the present invention is utilized for de novo sequencing of nucleic acids.

In order to illustrate what mismatch information or the pattern of mismatches may include, a hypothetical example will be described. Suppose the target nucleic acid is an 8-mer (meaning that the target nucleic acid consists of eight bases or nucleotides) and that the target is exposed to a chip including the complete set of 8-mer probes. In order to simplify this example, further assume that the 1 probe that is perfectly complementary to the target and the 24 probes that contain a single base mismatch (i.e., perfectly complementary except for a single base mismatch) have the highest hybridization intensities because they hybridize most strongly to the target.

Thus, if the target is ACTGGTCT-3', the following would be the probes having the highest measured intensities in this example:

Perfect complement TGACCAGA-5'

One base mismatches

GGACCAGA-5'
AGACCAGA-5'
CGACCAGA-5'

5

TAACCAGA-5'
TCACCAGA-5'
TTACCAGA-5'

10

.
. .
.

and so forth for the other six positions. The set of these 25 probes may be analyzed to sequence the target nucleic acid. Although typically, the target nucleic acid is longer than the probes, the example provides a good illustration of aspects of the present invention.

15

For many reasons, probes that are perfectly or exactly complementary to the target may not have the highest hybridization intensities. Therefore, a probe that is perfectly complementary to the target often cannot be identified from the rank order of hybridization intensities. The present invention utilizes mismatch information among the nucleic acid probes to sequence the target where all of the nucleic acid probes that are perfectly complementary to part of the target may not be readily identified.

20

25

The present invention identifies neighbor-rich probes which are then utilized to sequence the target nucleic acid. A "neighbor-rich probe" is a probe that is related to many other probes in the probe space by a single base mismatch. A probe that has a single base mismatch with another probe will be referred to as a "single base mismatch neighbor." Neighbor-rich probes may be identified according to mismatch information as follows.

30

35

After a set of probes is identified, each probe in the set is compared to the other probes to determine how the probe's sequence compares to the other probes. In the example above, one probe differs from the other 24 probes by a single base mismatch (i.e., $m = 1$, where m is the number of mismatches). Thus, this one probe is related to or has 24 single base mismatch neighbors.

40

By contrast, twenty-four probes differ from 3 other probes in the set by a single base mismatch and from 21 other

probes in the set by a double base mismatch (i.e., $m = 2$). In this simple example, the perfectly complementary probe may be identified as a neighbor-rich probe from the mismatch information because it has many single mismatch neighbor probes in the probe space. The perfectly complementary probe had 8 times as many single mismatch neighbor probes as nearly-complementary probes. Although the hybridization conditions did not allow identification of the perfectly complementary probes, an analysis of mismatch information may be utilized to identify the perfectly complementary probe. In practice, mismatch information may be utilized for de novo sequencing of a target nucleic acid where oligonucleotide probes are contacted with the target under conditions that do not allow differentiation between those probes that are perfectly complementary to part of the target and those probes that are not.

In this example, the sequence of the target was known. However, in many applications including de novo sequencing the sequence of the target is unknown. Nevertheless, the example is useful in demonstrating how neighbor-rich probes may be identified.

A. One Embodiment

Fig. 6 shows a high level flow of sequencing utilizing mismatch information. At step 500, hybridization intensities from probes or other data indicative of binding affinity are input into the system. The system may receive the hybridization intensities many different ways. The system may operate the scanning device directly, the system may receive the hybridization intensities from another computer system that measured the intensities, or an operator may manually enter the data. There may be thousands or tens of thousands of hybridization intensities that correspond to nucleic acid probes on a chip. Typically, the chip includes all possible probes of a specific length in order to sequence the target.

At step 502, the system selects a set of probes associated with the highest hybridization intensities (i.e., that show the strongest binding affinity). Selecting the

probes with the highest hybridization intensities may be done in any number of ways. For example, the system may use an intensity threshold value and select the probes whose hybridization intensities are higher than the intensity threshold (e.g., 100 photon counts). The system may select a specific number or percentage of probes (e.g., 50 probes or the top 10%) that have the highest hybridization intensities. Additionally, the system may select the probes that have a hybridization intensity greater than a specific percentage (e.g., 40%) of the highest hybridization intensity.

After the set of probes with the strongest binding affinity is selected, the system calculates the number of single base mismatch neighbors for each probe in the set at step 504. For example, in one embodiment, a probe is first selected in order to compare the selected probe to the other probes. The system then determines how many of the other probes in the set are identical to the selected probe except for a single base mismatch at one base position. The number of single base mismatch neighbors is calculated for each of the probes having the highest hybridization intensities. Additionally, the system may calculate and utilize the number of double base mismatches in an extension of the concepts herein.

At step 506, the system selects the probes in the set with the highest number of single base mismatch neighbors. Selecting the probes with the highest number of neighbors may be done in any number of ways including utilizing a threshold, a specific number of the probes, or greater than a specific percentage of the highest number of neighbors. In one embodiment, the system selects the probes in the set with the highest number of neighbors and the highest hybridization intensities. In other words, a second intensity threshold is utilized to further reduce the set of probes.

The selected probes with the highest number of single base mismatch neighbors are the neighbor-rich probes. The neighbor-rich probes are then aligned at step 508. The neighbor-rich probes are aligned or compiled so that they have the most bases in common. Thus, neighbor-rich probes that

have a single base mismatch are aligned to form an aligned set of probes. Aligned sets of probes are then aligned in a skewed fashion in the way that reduces the number of base mismatches between sets of probes. At step 510, the aligned probes are utilized to sequence the target nucleic acid sequence. The target may be sequenced in many different ways including the formation of a consensus sequence may be produced as described in the following example.

B. Example

A target of 5'-AGTTGTAGTGGATGG was exposed to a chip containing 8-mer probes. The highest hybridization intensity was 331 photon counts. An intensity threshold of 90 photon counts was utilized and there were 133 probes that had a hybridization intensity greater than the intensity threshold of 90. These 133 probes provided the set of probes with the highest hybridization intensities and are as follows:

probe	intensity	m = 1	m = 2
3'-ACATCACC	331	12	10
3'-CATCACCT	286	11	12
3'-ATCACCTA	323	10	8
3'-CATCACCA	253	8	12
3'-ACAACATC	331	7	10
3'-AATCACCT	131	7	14
3'-ACATCACA	330	7	10
3'-ACCTACCA	280	7	6
3'-CACCTACC	204	7	2
3'-ACTCACCT	188	7	11
3'-CCATCACC	270	6	13
3'-TTCACCTA	134	6	10
3'-ACACCACC	98	6	14
3'-TCAACATC	331	6	10
3'-TCATCACC	238	5	17
3'-CTCACCTA	203	5	10
3'-ACACCAAC	122	5	10
3'-ACACCTAC	272	5	8
3'-ACCTACCC	108	5	7
3'-TCATCACA	147	5	7
3'-CAACATCA	275	5	5
3'-CAACACCT	183	5	15
3'-CACCACCT	113	5	15
3'-ATCACCAC	157	5	6
3'-TCCACCTA	112	5	7
3'-TCACCTAC	248	5	6
3'-TGCACCTA	105	5	6
3'-TATCACCT	143	5	13
3'-CCACCTAC	208	5	5
3'-ACAACACC	147	5	18
3'-CCAACATC	325	4	10
3'-GCATCACC	262	4	15

	3'-GCACCTAC	199	4	6
	3'-AACATCAC	148	4	1
	3'-AGTCACCT	90	4	9
5	3'-CAACATCT	101	4	5
	3'-CCATCACA	181	4	9
	3'-CGCACCTA	127	4	6
	3'-TCAACACC	129	4	13
	3'-ATCACCTT	129	4	9
	3'-ACACCTAA	155	4	6
10	3'-CAACACCA	100	4	14
	3'-ACACACCT	305	4	11
	3'-TCACCTAA	176	4	4
	3'-ACACACCA	174	4	8
15	3'-CAGCACCT	111	4	13
	3'-ATCACCAA	115	4	11
	3'-ATCACCTC	137	4	13
	3'-GCATCACA	156	4	7
	3'-TACACCTA	96	4	7
20	3'-CCTCACCT	102	4	14
	3'-TCAACCTC	132	4	6
	3'-CACCACCA	91	4	13
	3'-CATCACCC	131	4	17
	3'-GCAACATC	319	4	9
25	3'-CATCAACC	105	4	6
	3'-CACCTACA	187	4	5
	3'-ACACCATC	128	4	13
	3'-ACCTACCT	112	4	11
	3'-ATTACCTT	91	4	10
30	3'-CCACCTAA	111	4	5
	3'-GCACCTAA	127	4	3
	3'-AGCACCTA	148	4	11
	3'-GTCACCTA	141	3	12
	3'-ACATCACT	164	3	13
35	3'-CATCACCG	163	3	15
	3'-CCCTACCA	133	3	6
	3'-ACCTACCG	119	3	8
	3'-ACAGCACC	101	3	14
	3'-ATCACCCA	106	3	12
40	3'-CGTCACCT	114	3	12
	3'-CAACATCC	148	3	7
	3'-ACAACCTC	114	3	10
	3'-ATCAACCT	120	3	7
	3'-ACCAACCA	104	3	12
45	3'-GCCTACCA	111	3	4
	3'-CACCAACA	119	3	8
	3'-ACTCACCA	143	3	12
	3'-ACCACCTA	141	3	17
	3'-CTATCACC	100	3	4
50	3'-CAACATCG	137	3	4
	3'-ACGCACCT	110	3	11
	3'-TCACCATC	102	3	10
	3'-CACACCTA	102	3	7
	3'-CACCTACT	106	3	6
55	3'-CACCAACC	103	3	11
	3'-ATCACCTG	106	3	9
	3'-ACATCACG	149	3	13
	3'-GCAACCTC	93	3	4

	3'-AAGCACCT	92	3	11
	3'-ATCATCAC	90	3	6
	3'-TCCTACCA	91	3	4
5	3'-ATCAACCA	103	3	5
	3'-ACCAACCT	97	3	10
	3'-GATCACCT	102	3	14
	3'-TCACCAAC	102	3	10
	3'-ACCTACTC	102	3	3
10	3'-CACCTACG	99	3	5
	3'-CCTCACCA	91	3	12
	3'-ATCACCAT	125	3	8
	3'-TCAACCTA	104	2	8
	3'-ACCATCAC	113	2	5
15	3'-CATCTACC	94	2	8
	3'-CAATCACC	94	2	6
	3'-ACATCAAC	154	2	15
	3'-ACCTACAC	113	2	5
	3'-ACACATCA	128	2	6
20	3'-CCACATCA	90	2	7
	3'-TATCACCA	97	2	12
	3'-CACATCAC	154	2	3
	3'-TCAACACA	97	2	8
	3'-TACCACCT	91	2	9
25	3'-ATCCACCT	105	2	11
	3'-ACACACCG	122	2	5
	3'-ACACCACA	90	2	14
	3'-ATATCACC	96	2	12
	3'-TACATCAC	128	2	2
30	3'-CAACCTAC	116	1	6
	3'-CATCACAA	107	1	7
	3'-ACCTCACC	102	1	13
	3'-ACCAACTC	96	1	9
	3'-TATCAACC	94	1	6
35	3'-TACCTACC	99	1	8
	3'-ACCACATC	128	1	12
	3'-ATCACAAC	153	1	5
	3'-CCTACATC	93	1	4
	3'-CACCTAAC	95	1	7
40	3'-CCTACCAA	128	0	0
	3'-TACACACC	91	0	2
	3'-CAACCATC	93	0	5
	3'-GTTAAGAG	329	0	0
	3'-AGCAACAT	94	0	3
45	3'-TCTATGCG	33	0	0

where the columns denoted $m = 1$ and $m = 2$ indicate the number of single and double base mismatch neighbors, respectively. Thus, each probe was compared to the other 132 probes to determine the number of single and double base mismatches the probe had with the other probes. The highest number of single base mismatch neighbor probes was 12 and the probes are presented in decreasing order according to the number of single base mismatch neighbor probes.

A set of neighbor-rich probes was identified by selecting the probes that had a hybridization intensity greater than 40% of 331 ($0.40 \times 331 = 132.4$), and the number of single base mismatch neighbors greater than 40% of 12 ($0.40 \times 12 = 4.8$). The following is the list of neighbor-rich probes selected in this manner:

probe	intensity	m = 1	m = 2
3'-ACACCTAC	272	5	8
3'-ACATCACC	331	12	10
3'-ACATCACA	330	7	10
3'-ACAACACC	147	5	18
3'-ACAACATC	331	7	10
3'-ACCTACCA	280	7	6
3'-ATCACCCAC	157	5	6
3'-ATCACCTA	323	10	8
3'-ACTCACCT	188	7	11
3'-CAACATCA	275	5	5
3'-CAACACCT	183	5	15
3'-CCACCTAC	208	5	5
3'-CCATCACC	270	6	13
3'-CTCACCTA	203	5	10
3'-CATCACCT	286	11	12
3'-CATCACCA	253	8	12
3'-TCACCTAC	248	5	6
3'-TCATCACC	238	5	17
3'-TCATCACA	147	5	7
3'-TCAACATC	331	6	10
3'-TTCACCTA	134	6	10
3'-TATCACCT	143	5	13

where again $m = 1$ and $m = 2$ indicates the number of single and double base mismatch neighbors, respectively, with other probes in the set of probes with a hybridization intensity greater than 90.

Once the neighbor-rich probes having a high hybridization intensity and high number of single base mismatch neighbors have been selected, the neighbor-rich probes were utilized to sequence the target nucleic acid sequence. The system utilized the frequency of bases at each position to produce a consensus sequence, where a "consensus sequence" is a sequence generated by neighbor-rich probes to sequence the target.

In order to produce a consensus sequence, the system aligned the neighbor-rich probes so that each probe had the highest number of bases in common with other probes. The following are the aligned neighbor-rich probes with the complement of the target sequence shown for reference. The

target sequence is known in this example but the target sequence may be an unknown sequence or only partially known.

```

5          A C A C C T A C
          A C A T C A C C
          A C A T C A C A
          A C A A C A C C
          A C A A C A T C
          A C C T A C C A
10         A T C A C C T A
          A T C A C C T
          A C T C A C T
          C A A C A C T
          C A A C A C
15         C C A T C A C C
          C T C A C C T A
          C A T C A C C A
          C A T C A C C A
          T C A C C A C
20         T C A T C C A
          T C A T C C A
          T C A T C C A
          T A T C A C C T
25 Target
    complement-T C A A C A T C A C C T A C C

```

After the neighbor-rich probes are aligned, the system counts the number or frequency of each base (A, C, G and T) at each position. After the frequency of bases at each position is calculated, the system produces a consensus sequence. In one embodiment, the base that occurred the most at a position is utilized to produce the consensus sequence if the base occurred more than 2 times and the frequency that the base occurred is greater than 50% at that position. The following is a matrix of base vs. frequency that was used to produce the consensus sequence in this manner:

```

Frequency
Base A 2 0 5 8 2 17 2 0 17 2 2 3 5 0 0
    C 0 5 0 1 15 4 3 22 2 16 16 1 1 2 1
    G 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
    T 1 0 0 2 1 1 18 1 1 1 0 9 0 0 0
Consensus - C A A C A T C A C C T A - -

```

Therefore, for the first position in the consensus sequence (left-most in the matrix), base A occurred 66% (2 divided by 3) of the time which is greater than 50%, however, the base did not occur more than 2 times so the base is called as ambiguous (i.e., "-").

The consensus sequence is the complement of the target; thus, the target is sequenced according to the complement of the consensus sequence. In this example, the target sequence was 5'-AGTTGTAGTGGATGG and it was correctly sequenced as 5'-GTTGTAGTGGAT (the terminal bases being ambiguous). The parameters for producing the consensus sequence may be varied according to the experimental data. For example, if the consensus sequence was formed solely by the bases that occur most often at each position, the consensus sequence would be perfectly complementary to the target nucleic acid for this data. However, this will not always be the case.

Software Appendix A provides a BASIC source code illustration of this embodiment of the invention. The source code is written in Quick BASIC for an IBM compatible personal computer.

C. Alternate Embodiment

Fig. 7 shows a high level flow of another embodiment of sequencing utilizing mismatch information. At step 550, hybridization intensities are input into the system. The system may receive the hybridization intensities many different ways. The system may operate the scanning device directly, the system may receive the hybridization intensities from another computer system that measured the intensities, or an operator may manually enter the data.

At step 552, the system selects a set of probes associated with the highest hybridization intensities. Selecting the probes with the highest hybridization intensities may be done in any number of ways. For example, the system may use an intensity threshold value and select the probes whose hybridization intensities are higher than the intensity threshold. The system may select a specific number or percentage of probes that have the highest hybridization intensities. Alternatively, the system may select the probes that have a hybridization greater than a specific percentage of the highest hybridization intensity.

After the set of probes with the highest hybridization intensities is selected, the system calculates

the number of straight mismatches for each probe in the set at step 554. "Straight mismatches" refers to base mismatches between probes where the bases at corresponding positions are compared (e.g., as was done in the previous embodiment). For example, a probe 3'-AACAT is compared to a probe 3'-AACTT by aligning the probes as follows:

```

3'-AACAT
3'-AACTT

```

Each probe has n bases, where $n = 5$. If the base at the 3' end is at position 1, there is a single mismatch at the fourth position where the A and T do not match. Thus, straight mismatches are determined by comparing bases at the same position in each probe (i.e., $z = 0$, where z indicates the number of bases one of the probes was skewed or offset from the other probe before comparing).

In one embodiment, a matrix is formed to show the straight mismatches between probes. Fig. 8 shows a straight mismatch matrix for 50 probes. For simplicity, each probe is assigned an identification number from 1 to 50. The numbers on the rows and columns of the matrix correspond to the identification number of the probe. The value in the matrix represents the number of straight mismatches between the probes designated by the row and column. If there are more than 2 mismatches, the matrix shows a "." at the appropriate matrix location. Since a diagonal 600 of the matrix shows the number of straight mismatches between the same probe, the diagonal contains 0's because the probe is being compared to itself (i.e., there are no straight mismatches). Also, the matrix is identical on each side of diagonal 600.

The system calculates the number of skewed mismatches for each probe in the set at step 556. "Skewed mismatches" refers to base mismatches between probes where the bases at offset positions are compared. For example, a probe 3'-CGAATCAA is compared to a probe 3'-GCATCAAC by aligning the probes as follows:

```

3'-CGAATCAA
3'-GCATCCAC

```

Each probe has n bases, where $n = 8$. If the base at the 3' end is at position 1, bases at position 1 through 7 (or $n-1$)

of the first probe is compared to bases at position 2 through 8 (or n). As shown, there are two mismatches (double mismatch) when the probes are skewed a single base position (i.e., $z = 1$, where z indicates the number of bases one of the probes is skewed or offset from the other probe before comparing).

In one embodiment, a matrix is formed to show the skewed mismatches between probes. Fig. 9 shows a skewed mismatch matrix for 50 probes. As in Fig. 8, the rows and columns of the matrix correspond to the identification number of the probe, which is 1 to 50. The value in the matrix represents the number of skewed mismatches between the probes designated by the row and column. If there are more than 2 mismatches, the matrix shows a "." at the appropriate matrix location. As shown, a diagonal will not contain 0's and the matrix is not identical on each side of the diagonal. Although the probes were skewed a single base position, the probes may be skewed more positions when they are compared in an extension of the principles herein.

At step 558, the system selects the probes with less than some small number straight mismatches and less than some small number of skewed mismatches. In one embodiment, the system identifies the probes in the matrices that have less than 2 straight mismatches and less than 3 skewed mismatches. The parameters for selecting these probes with few mismatches may be varied according to the experimental data.

The selected probes are then aligned at step 560. The probes are aligned so that they have the most bases in common. The mismatch information concerning the straight mismatches and skewed mismatches is utilized to align the probes so that the number of mismatches between the probes is reduced. At step 562, the aligned probes are utilized to sequence the target nucleic acid sequence. The target may be sequenced in many different ways. For example, a consensus sequence may be produced as described in the following example.

D. Example

A target of 5'-AGTTGTAGTGGATGGT was exposed to a chip containing 10-mer probes. Fifty probes were selected that have the highest hybridization intensities (step 552). Figs. 11 and 12 show the straight and skewed mismatch matrices for the fifty probes (steps 554 and 556). Forty-seven probes were selected that have less than 2 straight mismatches with at least one other probe and less than 3 skewed mismatches with at least one other probe (step 558).

The straight and skewed mismatch information was utilized to align the 47 probes (step 560). For example, Fig. 9 shows that the probe identified as 2 on the row had 0 skewed mismatches with the probe identified as 1 on the column. Therefore, probes 2 and 1 align well if they are offset a single base position. The following are the aligned 47 probes:

```

AACATCACCT
CAACATCAC
ACATCACCTA
ACAACATCAC
CAACATCACA
ATCACCTACC
AACATCACCA
CACATCACCT
AACATCACCG
TCAACATCAC
CATCACCTAC
CCAACATCAC
ACACCTACCA
CAACATCACG
GCAACATCAC
ACATCACCTT
AACATCACCC
AGCACCTACC
AAACATCAC
ACCATCACCT
ACATCACCAT
CACCTACCAA
ACATCACCTC
ATCACCTACA
ACATCACCTG
ACACATCAC
GAACATCAC
TACATCACCT
TAACATCAC
CACCTACCAG
GACATCACCT
CACATCACCA
ATCATCACCT
ACCTACCATC
ACAACATCAA

```

CAACATCACT
 ACATCACCAA
 CACATCACCG
 ACATCACCTT
 CACCTACCAC
 CATCACCTAA
 TCACCTACCA
 CACCTACCAT
 CCATCACCTA
 ACATCACCCA
 ACATCACCGA
 ATCAACATCA

After the selected probes were aligned, the system counts the number of occurrences of each base (A, C, G and T) at each position. After the frequency of bases at each position is calculated, the system produces a consensus sequence which should be complementary to the target sequence. If the system utilized bases that occurred more than 2 times and the frequency that the base occurred is greater than 50% at that position, the consensus sequence 3'-CAACATCACCTACCA is produced.

The consensus sequence is ideally the complement of the target; thus, the target is sequenced according to the complement of the consensus sequence (consensus' where the prime denotes the complement). In this example, the target and consensus' sequence were as follows:

Target	AGTTGTAGTGGATGGT
Consensus'	GTTGTAGTGGATGGT

(one terminal base of the consensus sequence being ambiguous). Thus, the target was sequenced with a high degree of accuracy utilizing mismatch information. The parameters for producing the consensus sequence may be varied according to the experimental data.

Software Appendix B provides a BASIC source code illustration of this embodiment of the invention. The source code is written in Quick BASIC for an IBM compatible personal computer.

Conclusion

The above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this disclosure. Merely by way of example, while the invention is

illustrated with particular reference to the evaluation of DNA (natural or unnatural), the methods can be used in the analysis from chips with other materials synthesized thereon, such as RNA. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

5

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2

APPENDIX A

1001453.121401
1001453.121401

```
'SEARCH.BAS 11/21/94
```

```
'Finds pixels with intensities greater than fmin
'Calculates the match score of bright pixels
```

```
'The input file eightmer.dat is based on fs8mer.exe
'The output file score.dat lists the bright pixels
' and gives the number of 1 and 2 mismatch-related
' pixels
```

```
DIM a$(1000), f$(1000), m1$(1000), m2$(1000)
```

```
tstart = TIMER
inf$ = "eightmer.dat"
outf$ = "score.dat"
OPEN inf$ FOR INPUT AS #1
OPEN outf$ FOR OUTPUT AS #2
CLS
```

```
'Read the input file and store the bright pixels
```

```
fmin = 90 'threshold value for inclusion
```

```
n = 0: fmax = 0
```

```
WHILE NOT EOF(1)
```

```
LINE INPUT #1, g$
```

```
seq$ = MID$(g$, 1, 8)
```

```
intens = VAL(MID$(g$, 9, 6))
```

```
IF intens >= 90 THEN
```

```
    n = n + 1
```

```
    a$(n) = seq$
```

```
    f$(n) = intens
```

```
    IF intens > fmax THEN fmax = intens
```

```
PRINT n;
```

```
END IF
```

```
WEND
```

```
PRINT
```

```
PRINT USING "#### intensity values above ####"; n; fmin
```

```
PRINT USING "Highest intensity is ####"; fmax
```

```
'Calculate m1 and m2 for each bright pixel
```

```
' m1 and m2 are the number of other pixels that are related
```

```
' by 1 and 2 mismatches, respectively
```

```
m1max = 0 'Keep track of highest m1 score
```

```
FOR j = 1 TO n
```

```
PRINT j;
```

```
FOR i = 1 TO n
```

```
m = 0
```

```
FOR k = 1 TO 8
```

```
IF MID$(a$(j), k, 1) <> MID$(a$(i), k, 1) THEN m = m + 1
```

```
NEXT k
```

```
IF m = 1 THEN m1$(j) = m1$(j) + 1
```

```
IF m = 2 THEN m2$(j) = m2$(j) + 1
```

```
NEXT i
```

```
IF m1$(j) > m1max THEN m1max = m1$(j)
```

```
NEXT j
```

```
PRINT #2, USING "SEARCH.BAS & "; DATE$; TIME$
```

```
PRINT #2, USING "Input file: & Output file: &"; inf$; outf$
```

```
PRINT #2, USING "#### intensity values above ####"; n; fmin
```

```
PRINT #2, USING "Highest intensity is ####"; fmax
```

```

PRINT #2, USING "Greatest number of 1-mismatch relations is ##"; m1max
PRINT #2,
PRINT #2, "List of probes with highest intensity and best matching"
PRINT #2, " f m1 m2 sequence"
FOR k = 1 TO n
IF f%(k) > .4 * fmax AND m1%(k) > .4 * m1max THEN
PRINT #2, USING "#### ## ## &"; f%(k); m1%(k); m2%(k); a$(k)
END IF
NEXT k
PRINT #2, CHR$(12)

'Sort according to f
s% = n \ 2
DO WHILE s% > 0
FOR i% = s% TO n - 1
j% = i% - s% + 1
FOR j% = (i% - s% + 1) TO 1 STEP -s%
IF f%(j%) >= f%(j% + s%) THEN EXIT FOR
SWAP f%(j%), f%(j% + s%)
SWAP m1%(j%), m1%(j% + s%)
SWAP m2%(j%), m2%(j% + s%)
SWAP a$(j%), a$(j% + s%)
NEXT j%
NEXT i%
s% = s% \ 2
LOOP
PRINT #2,
PRINT #2, " f m1 m2 sequence"
FOR k = 1 TO n
PRINT #2, USING "#### ## ## &"; f%(k); m1%(k); m2%(k); a$(k)
NEXT k
PRINT CHR$(12)
'Sort according to m1
s% = n \ 2
DO WHILE s% > 0
FOR i% = s% TO n - 1
j% = i% - s% + 1
FOR j% = (i% - s% + 1) TO 1 STEP -s%
IF m1%(j%) >= m1%(j% + s%) THEN EXIT FOR
SWAP f%(j%), f%(j% + s%)
SWAP m1%(j%), m1%(j% + s%)
SWAP m2%(j%), m2%(j% + s%)
SWAP a$(j%), a$(j% + s%)
NEXT j%
NEXT i%
s% = s% \ 2
LOOP
PRINT #2,
PRINT #2, " f m1 m2 sequence"
FOR k = 1 TO n
PRINT #2, USING "#### ## ## &"; f%(k); m1%(k); m2%(k); a$(k)
NEXT k

PRINT USING "Time= ###.## seconds"; TIMER - tstart

```

APPENDIX B

1004453:124104

```

'CONSENS: BAS 1/8/95
'Derive a consensus sequence from the highest scoring probes

DIM a$(70), m$(1, 70, 70), f(70), s(-20 TO 20, 4)
CLS
INPUT "Input file: ", inf$
INPUT "Output file: ", outf$
OPEN inf$ FOR INPUT AS #1
OPEN outf$ FOR OUTPUT AS #2

LINE INPUT #1, descr$ 'File description
INPUT #1, pl 'Probe length
INPUT #1, n 'Number of sequences
FOR j = 1 TO n
LINE INPUT #1, a$(j)
NEXT j
CLOSE #1

'Initialize the mismatch matrix
FOR z = 0 TO 1: FOR i = 1 TO n: FOR j = 1 TO n
m$(z, i, j) = 100
NEXT j: NEXT i: NEXT z

PRINT #2,
PRINT #2, "CONSENS.BAS "; DATE$; " "; TIME$
PRINT #2, : PRINT #2,
PRINT #2, "Input file: "; inf$; " Output file: "; outf$
PRINT #2, descr$
PRINT #2, USING "The ## ##-mer sequences with the highest scores are: ";
n; pl
PRINT #2,
FOR j = 1 TO n
PRINT #2, USING "## "; j; a$(j)
NEXT j
PRINT #2, : PRINT #2,

z = 0
PRINT #2, USING "z=##"; z
PRINT #2, " ";
FOR k = 1 TO n: PRINT #2, USING "##"; k; : NEXT k
FOR i = 1 TO n
PRINT #2,
PRINT #2, USING "## "; i;
FOR j = 1 TO n
m = 0
FOR k = 1 TO pl
IF MID$(a$(j), k, 1) <> MID$(a$(i), k, 1) THEN m = m + 1
NEXT k
m$(0, i, j) = m
IF m <= 2 THEN PRINT #2, USING " ##"; m; ELSE PRINT #2, " .";
NEXT j
NEXT i
PRINT #2, : PRINT #2,

z = 1
PRINT #2, USING "z=##";
PRINT #2, " ";

```

```

FOR k = 1 TO n: PRINT #2, USING "##"; k; : NEXT k
FOR i = 1 TO n
  PRINT #2,
  PRINT #2, USING "## "; i;
  FOR j = 1 TO n
    m = 0
    FOR k = 1 TO pl - 1
      IF MID$(a$(j), k, 1) <> MID$(a$(i), k + 1, 1) THEN m = m + 1
    NEXT k
    m%(1, i, j) = m
    IF m <= 2 THEN PRINT #2, USING " ##"; m; ELSE PRINT #2, " .";
  NEXT j
NEXT i

PRINT #2, : PRINT #2,

'Mark all sequences with a 100 tag
FOR i = 1 TO n: f(i) = 100: NEXT i
'Designate the first sequence as the origin
f(1) = 0

'Find the frames of sequences that can be aligned
FOR i = 1 TO n
  FOR j = 1 TO n
    IF m%(1, i, j) <= 2 AND f(i) <> 100 THEN
      f(j) = f(i) + 1
    END IF
  NEXT j
NEXT i

FOR i = 1 TO n
  FOR j = 1 TO n
    IF m%(1, j, i) <= 2 AND f(i) <> 100 THEN
      f(j) = f(i) - 1
    END IF
  NEXT j
NEXT i

FOR i = 1 TO n
  FOR j = i + 1 TO n
    IF m%(0, i, j) <= 1 AND f(i) <> 100 THEN
      f(j) = f(i)
    END IF
  NEXT j
NEXT i

PRINT #2, : PRINT #2,
PRINT #2, "Alignment criteria: <=1 mismatch allowed for z=0"
PRINT #2, "      <=2 mismatches for z=1"

PRINT #2,
PRINT #2, "The aligned sequences are:"
'Print the aligned sequences
FOR i = 1 TO n
  IF f(i) <> 100 THEN
    PRINT #2, SPACE$(15 + f(i)); a$(i)
  END IF

```

1004953, 121101

```

NEXT i
PRINT #2, : PRINT #2,

'Accumulate the sequence scores
offset = 0
FOR i = 1 TO n
IF f(i) <> 100 THEN
    FOR k = 1 TO pl
        g = INSTR("ACGT", MID$(a$(i), k, 1))
        s(offset + k + f(i), g) = s(offset + k + f(i), g) + 1
        'PRINT offset + k + f(i); g; "      ";
    NEXT k
END IF
NEXT i

PRINT #2, CHR$(12)
PRINT #2, "CONSENS.BAS "; DATE$; " "; TIME$
PRINT #2, USING "Input file: &      Output file: &"; inf$; outf$
PRINT #2, USING "### ##mer sequences"; n; pl
PRINT #2, descr$
PRINT #2,

PRINT #2, "The frequencies of bases in the aligned sequences are:"
PRINT #2,
'Print the scores
FOR g = 1 TO 4
FOR j = -10 TO 18
PRINT #2, USING "## "; s(j, g);
'PRINT USING "## "; s(j, g);
NEXT j
PRINT #2,
NEXT g

'Find and print the consensus
c$(0) = "-": c$(1) = "A": c$(2) = "C": c$(3) = "G": c$(4) = "T"
FOR j = -10 TO 18
most = 0: mg = 0: sum = 0: b$ = "-"
FOR g = 1 TO 4
IF s(j, g) > most THEN most = s(j, g): mg = g
sum = sum + s(j, g)
NEXT g

'A base is defined if present in at least 2 sequences
'  and 55% of those aligned at that position

IF most >= 3 THEN
    IF most / sum > .5 THEN b$ = c$(mg)
END IF
PRINT #2, USING " & "; b$;
cons$ = cons$ + b$
NEXT j
PRINT #2, : PRINT #2, : PRINT #2, "The consensus sequence is: "; cons$:
PRINT #2,
PRINT cons$

PRINT #2, : PRINT #2,
PRINT #2, "The correct sequence is TCAACATCACCTACCA"

```

